[illegible] FILE COPY

(2)

SECURITY CLASSIFICATION OF THIS PAGE

AD-A221 946

DTIC ELECTE MAY 0 4 1990 S D D

# REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a.

1b. RESTRICTIVE MARKINGS

2a.

3. DISTRIBUTION/AVAILABILITY OF REPORT
Approved for Public Release;
Distribution Unlimited

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFOSR-TR- 90-0396

6a. NAME OF PERFORMING ORGANIZATION
Bowling Green State University
Department of Mathematics & Stats

6b. OFFICE SYMBOL (If applicable)

7a. NAME OF MONITORING ORGANIZATION
Air Force Office of Scientific Research

6c. ADDRESS (City, State, and ZIP Code)
Bowling Green, OH 43403-0221

7b. ADDRESS (City, State, and ZIP Code)
Bldg 410
Bolling AFB, DC 20332-6448

8a. NAME OF FUNDING/SPONSORING ORGANIZATION Air Force
Office of Scientific Research

8b. OFFICE SYMBOL (If applicable)
NM

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
AFOSR-88-0234

8c. ADDRESS (City, State, and ZIP Code)
Bldg 410
Bolling AFB, DC 20332

10. SOURCE OF FUNDING NUMBERS

| PROGRAM ELEMENT NO. | PROJECT NO. | TASK NO | WORK UNIT ACCESSION NO. |
|---|---|---|---|
| 61102F | 2304 | A3 | |

11. TITLE (Include Security Classification)
Computational Methods for Problems in Fluid Dynamics (Unclassified)

12. PERSONAL AUTHOR(S)
So-Hsiang Chou

13a. TYPE OF REPORT
Final

13b. TIME COVERED
FROM 7/1/88 TO 12/31/89

14. DATE OF REPORT (Year, Month, Day)
1989, February

15. PAGE COUNT
30

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

| FIELD | GROUP | SUB-GROUP |
|---|---|---|
| | | |
| | | |

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Reduced Basis, Projection Method, Linear Solver, Sparse Symmetric Linear Systems, Contraction Number, Hybrid Difference Methods

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The research reported herein addresses the reduced basis method, which is based on the combination of a basic iterative method and a projection method. The convergence is analyzed and some error bounds are established. The relationship between the reduced basis method and the preconditioned conjugate gradient method is discussed. Also included is a practical implementation of the reduced basis method when a pseudoresidual-based Krylov space is chosen. The final section of this report dwells on the weight selection procedures of hybrid difference methods for the linear convective equation. The procedures are based on the ability of hybrid difference methods to conserve the discrete weighted energy.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
David A. Nelson, Lt. Col.

22b. TELEPHONE (Include Area Code)
(202)767-5025

22c. OFFICE SYMBOL
AFOSR/NM

Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

# COMPUTATIONAL METHODS FOR PROBLEMS IN FLUID DYNAMICS

So-Hsiang Chou

BGSU Technical Report No. 90-03

## FINAL TECHNICAL REPORT

FOR THE PERIOD 1 JULY 1988 THROUGH 31 DECEMBER 1989
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

GRANT NO. AFOSR-88-0234

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

DEPARTMENT OF MATHEMATICS AND STATISTICS
BOWLING GREEN STATE UNIVERSITY
BOWLING GREEN, OHIO 43403-0221

Accesion For
NTIS CRA&I ☑
DTIC TAB □
Unannounced □
Justification ________

By ________
Distribution /

Availability Codes

Dist | Avail and / or Special
A-1

DTIC COPY INSPECTED

90 06 [illegible] 199

## 0. Introduction

The approximation of partial differential equations by the finite element or finite difference methods often leads to large sparse linear systems. Traditionally, one employs basic iterative methods to solve them. However, basic iterative methods exhibits slow convergence if the meshsize is small. This is due to the fact that basic iterative methods can not damp out low frequency modes of the errors. To remedy this disadvantage, multigrid methods combine basic iterative methods with other methods that are complementary. One of the reasons that accounts for the effectiveness of multigrid methods seems to be the idea of approximating the solution of a large system from a subspace whose dimension is small. Finding such a subspace is by no means an easy task. In this report we investigate the residual based reduced basis method (RRBM) for solving large sparse symmetric positive definite systems. In a sense, it belongs to the class of two-grid methods, although no geometric grids are involved. The outline of this report is as follows. Section 1 is devoted to the convergence behavior of general projection processes. A generic bound is provided for a class of pseudoresidual-based projection processes. In Section 2, we give mild conditions that ensure the convergence of general additive correction procedures. In Section 3, we describe the RRBM and prove a convergence theorem for it. Section 4 is devoted to the practical implementation of the RRBM. In Section 5 we relate the RRBM to the preconditioned conjugate gradient method under a certain condition. It is shown that the RRBM is more flexible than the preconditioned conjugate gradient method. In Section 6 we extend some of the previous results to the case of nonsymmetric linear systems having positive definite symmetric part. A rather general convergence analysis is given there. Finally, Section 7 is devoted to the weight selection principle for the linear convective equation in $R^n$. The principle is based on the ability of hybrid difference methods to conserve the discrete weighted energy.

## 1. Additive Correction Scheme

Let $A$ be a symmetric positive definite (SPD) matrix of order $n$. Consider the linear system of equations

$$Au = b. \tag{1.1}$$

The general additive correction scheme for solving (1.1) can be described as follows.

<u>ALGORITHM 1</u>: *Let* $u^{(0)}$ *be given. Choose a fixed integer* $\upsilon$. *For* $k = 1,2,\ldots,$ *do*

*Step 1. Initialization:*

$$w_k^{(0)} = u^{(k)}.$$

*Step 2. Presmoothing:*

$$w_k^{(j)} = Gw_k^{(j-1)} + Q^{-1}b, \quad j = 1,2,\ldots,\upsilon. \tag{1.2a}$$

*Step 3. Defect Computation:*

$$d_k^{(\upsilon)} \equiv b - Aw_k^{(\upsilon)} = A(u - w_k^{(\upsilon)}) \equiv A\varepsilon_k^{(\upsilon)}. \tag{1.2b}$$

*Step 4. Additive Correction:*

$$u^{(k+1)} = w_k^{(\upsilon)} + \varepsilon_R, \tag{1.2c}$$

*where* $\varepsilon_R$ *is a computationally inexpensive approximation of* $\varepsilon_k^{(\upsilon)}$.

Some remarks are in order. Steps 1 and 2 in the above algorithm consists of the usual basic iterative method with the iteration matrix $G$ and the splitting matrix $Q$. It is well known that basic iterative methods applied to the linear systems arising from discretization of partial differential equations are not effective once the high frequency modes have been damped out. Hence one stops after a certain number of steps of the basic iterative method. At this juncture, tne error $\varepsilon_k^{(\upsilon)}$ will lie mainly in the subspace of low frequency modes. Intuitively, this means the error $\varepsilon_k^{(\upsilon)}$ can be well approximated oy solving an approximating system of the linear system (1.2b) in a smaller subspace. The solution $\varepsilon_R$ (R for "reduced") is then added to $w_k^{(\upsilon)}$. The success of Algorithm 1 depends whether one can judiciously choose the above-mentioned subspace. In this report we confine ourselves to projection processes in which $\varepsilon_R$ is the projection of $\varepsilon_k^{(\upsilon)}$ onto a certain subspace. We now study the effects of steps 3 and 4.

Let $w$ be an approximation of the solution $u$ to the system (1.1). Let $S = [s_1, s_2, \ldots, s_m]$ be a full rank matrix consisting of column vectors $s_i$, $1 \leq i \leq m$. The following describes how to get a new approximation $\tilde{u}$ from $w$.

Galerkin Approximation

$$\begin{cases} d \equiv b - Aw = Au - Aw \equiv Ae. & (1.3a) \\ \text{Find } \varepsilon_R \in Rg(S) \equiv \text{the column space of } S \text{ such that} & \\ S^T(d - A\varepsilon_R) = 0. & (1.3b) \\ \tilde{u} \equiv w + \varepsilon_R. & (1.3c) \end{cases}$$

The above is equivalent to

$$\varepsilon_R = S(S^TAS)^{-1}S^TAe, \qquad (1.4a)$$

$$\tilde{e} \equiv u - \tilde{u} = e - \varepsilon_R = e - S(S^TAS)^{-1}S^TAe. \qquad (1.4b)$$

Let $(x,y)_E = x^TAy$, the energy product associated with the matrix $A$. It is easy to see that $P_S \equiv S(S^TAS)^{-1}S^TA$ is the orthogonal projector onto $Rg(S)$. Let $\|\cdot\|_E$ be the energy norm induced by $(\cdot,\cdot)_E$. Then from (1.4b) we have

$$\|\tilde{e}\|_E = \|(I - P_S)e\|_E = \min_{Z \in Rg(S)} \|e - z\|_E \qquad (1.5a)$$

or

$$\|\tilde{u} - u\|_E = \min_{v \in w + Rg(S)} \|v - u\|. \qquad (1.5b)$$

It is known that if $Rg(S) = K^{(m)}(d,A) \equiv \mathrm{Span}\{d, Ad, \ldots, A^{(m-1)}d\}$, then (1.5a) results in a bound involving a Chebyshev polynomial of the first kind. That is

$$\|\tilde{e}\|_E \le \frac{1}{T_m\left(\frac{b-a}{b+a}\right)} \|e\|_E,$$

where $T_m$ is the Chebyshev polynomial of degree $m$ and $[a,b]$ is the smallest interval containing the spectrum of the matrix $A$. Note that the conjugate gradient method falls into this category. We prove below that a generic inequality exists for all projection processes (1.4) if $d \in Rg(S)$.

LEMMA 1.1 [Lemma 3.1, D. Braess, 1981] *Let the Hilbert space* $U$ *be a direct sum of its subspaces* $V$ *and* $W$. *Assume that there is a* $\gamma < 1$ *such that a strengthened Cauchy inequality holds:*

$$|(v,w)| \le \gamma\|v\|\|w\|, \quad v \in V, \quad w \in W.$$

*If* $P_W u = 0$, *then*

$$\|u - P_V u\| \le \gamma\|u\|.$$

*Here* $P_W$ *and* $P_V$ *are orthogonal projectors to* $W$ *and* $V$, *respectively.*

THEOREM 1.1. *Let* $\tilde{e}$ *and* $e$ *be as in (1.4). Assume that the defect* $d \in Rg(S)$. *Then*

$$\|\tilde{e}\|_E \le \frac{k(A) - 1}{k(A) + 1} \|e\|_E, \tag{1.6}$$

*where* $k(A) = \|A\|_2 \|A^{-1}\|_2$ *and* $\|\cdot\|_2 \equiv$ the 2-norm of $A$.

*Proof.* We set various spaces corresponding to Lemma 1.1 as follows.

$U = R^n$ endowed with the energy product,

$V = Rg(S)$, $W = Rg^{\perp}(S)$ = the orthogonal complement

of $V$ with respect to the Euclidean inner product.

Obviously,

$$R^n = V \oplus W.$$

Furthermore, $P_W e = 0$ since $Ae = d$. Also, $P_V \equiv S(S^TAS)^{-1}S^TA$. The constant $\gamma$ in Lemma 1.1 can be determined as follows. For $v \in V$ and $w \in W$,

$$
\begin{aligned}
(v,w)_E &= v^T A w \\
&= v^T A w - \alpha v^T w \qquad \forall \alpha > 0 \\
&= v^T (A - \alpha I) w \\
&= (A^{1/2} v)^T [A^{-1/2} (A - \alpha I) A^{-1/2}) A^{1/2} w.
\end{aligned}
$$

Hence

$$\left|(v,w)_E\right| \le \|v\|_E \, \|w\|_E \, \|I - \alpha A^{-1}\|_2.$$

Let $0 < \lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$ be the eigenvalues of $A$. Now

$$
\begin{aligned}
\gamma &\equiv \min_{\alpha>0} \|I - \alpha A^{-1}\|_2 \\
&= \min_{\alpha>0} \max_{\lambda \in \sigma(A^{-1})} \left|1 - \alpha\lambda\right|, \quad \sigma(A^{-1}) = \text{the spectrum of } A^{-1} \\
&= \min_{\alpha>0} \{\left|1 - \alpha\lambda_1^{-1}\right|, \ldots, \left|1 - \alpha\lambda_n^{-1}\right|\} \\
&= \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} \\
&= (k(A) - 1)/(k(A) + 1).
\end{aligned}
$$

The conclusion of the theorem then follows easily by noting that

$$\tilde{e} = (I - P_V)e.$$

QED.

We remark that Theorem 1.1 covers the steepest descent method and the conjugate gradient method. Inequality (1.6) for the steepest descent method is usually derived from Kantorovich inequality (see, e.g. Luenberger, 1984).

Theorem 1.1 can be generalized as follows.

<u>THEOREM 1.2</u>. *Let* $\tilde{e}$ *and* $e$ *be as in (1.4). Assume that* $d \in Rg(S)$. *Let* $C$ *be a SPD matrix such that* $C(Rg(S)) \subset Rg(S)$. *Then*

$$\|\tilde{e}\|_E \le \frac{k(C^{-1/2}AC^{-1/2}) - 1}{k(C^{-1/2}AC^{-1/2}) + 1} \|e\|_E.$$

*Proof.* The corresponding subspaces are the same as in Theorem 1.1. The only detail changed is the choice of $\gamma$. Let $v \in V$ and $w \in W$. Then

$$\begin{aligned}(v,w)_E &= w^T A v \\ &= w^T A v - \alpha w^T C v \qquad \forall \alpha > 0 \\ &= w^T (A - \alpha C).\end{aligned}$$

As before, we have

$$|(v,w)_E| \le \|w\|_E \, \|v\|_E \, \|I - \alpha A^{-1/2} C A^{-1/2}\|_2.$$

Set $B^{-1} \equiv A^{-1/2} C A^{-1/2}$. For a SPD matrix $D$ we denote the eigenvalues of $D$ as $0 < \lambda_1(D) \le \lambda_2(D) \le \ldots \le \lambda_n(D)$. Define the constant $\gamma$ in Lemma 1.1 as

$$\gamma = \frac{k(B) - 1}{k(B) + 1} = \frac{\lambda_n(B) - \lambda_1(B)}{\lambda_n(B) + \lambda_1(B)}.$$

But

$$\begin{aligned}\lambda_n(B) &= \rho(B) = \text{the spectral radius of } B \\ &= \rho(A^{1/2} C^{-1} A^{1/2}) \\ &= \rho(C^{-1} A) \\ &= \rho(C^{-1/2} C^{-1/2} A) \\ &= \rho(C^{-1/2} A C^{-1/2}),\end{aligned}$$

and

$$\begin{aligned}\lambda_1(B) &= 1/(1/\lambda_1(B)) \\ &= 1/\rho(B^{-1}) \\ &= 1/\rho(A^{-1/2} C A^{-1/2}) \\ &= 1/\rho(C A^{-1}) \\ &= 1/\rho(C^{1/2} A^{-1} C^{1/2}) \\ &= 1/\rho((C^{-1/2} A C^{-1/2})^{-1}) \\ &= \lambda_1(C^{-1/2} A C^{-1/2}).\end{aligned}$$

QED.

**THEOREM 1.3.** *Let* $S = Cd \in R^{n \times 1}$ *and* $0 \ne d$. *If* $C$ *is SPD then*

$$\|\tilde{e}\|_E \le \frac{k(C^{1/2} A C^{1/2}) - 1}{k(C^{1/2} A C^{1/2}) + 1} \|e\|_E.$$

*Proof.* In reference to Lemma 1.1, we set

$$V = \text{Span}\{Cd\},$$

$$W = \{w \mid w^T d = 0\} \qquad ,$$
$$U = R^n.$$

It is easy to see that

$$U = V \oplus W,$$

and

$$P_W e = 0.$$

Let $v \in V$ and $w \in W$. $v \in V$ implies that $v = \beta Cd$ for some $\beta$. Thus

$$\begin{aligned}(v,w)_E &= v^T A w \\ &= v^T A w - \alpha\beta d^T w, \qquad \forall\alpha > 0 \\ &= v^T A w - \alpha v^T (C^{-1})^T w \\ &= v^T (A - \alpha C^{-1}) w.\end{aligned}$$

Now the rest of proof is just as in Theorem 1.2, we omit the details.

QED.

If we take $C = I$, the identity matrix, we recover the steepest descent method. If $A$ has Property A, then we can take $C = D^{-1}$, where $D$ is the diagonal part of the matrix $A$. It is well known in this case that $D^{-1/2}AD^{-1/2}$ has smaller condition (see, e.g. Young 1971, p. 214). In general, we can take $C = Q^{-1}$, where $Q$ is the splitting matrix of a basic iterative method applied to the system (1.1).

## 2. Convergence Analysis

We now turn to convergence analysis of Algorithm 1. Since we are interested in the relationship between two consecutive iterates, we shall drop the subindex $k$ appearing in (1.2a-1.2c). One cycle of Algorithm 1 with Galerkin approximation is as follows.

Given $u^{(k)}$.

1. Initialization:

$$w^{(0)} = u^{(k)}.$$

2. Presmoothing:

$$w^{(j)} = Gw^{(j-1)} + Q^{-1}b, \quad j = 1,2,\ldots,\upsilon. \qquad (2.1a)$$

3. Defect Computation:

$$d \equiv b - Aw^{(v)} = A(u - w^{(v)}) \equiv A\varepsilon^{(v)}. \tag{2.1b}$$

4. Additive Co^rrection:

Let a full rank matrix $S \in R^{n \times m}$ be given.

(4.a) Solve

$$S^T ASz = S^T d. \tag{2.2}$$

(4.b) Set

$$\varepsilon_R = Sz. \tag{2.3}$$

(4.c) Set

$$u^{(k+1)} = w^{(v)} + \varepsilon_R. \tag{2.4}$$

Define the k-th error vector as

$$e^{(k)} \equiv u - u^{(k)}. \tag{2.5}$$

It is easy to see that

$$\varepsilon^{(v)} = G^v e^{(k)} \tag{2.6}$$

and

$$e^{(k+1)} = \varepsilon^{(v)} - \varepsilon_R. \tag{2.7}$$

From (1.4b) with $\tilde{e} = e^{(k+1)}$ and $e = \varepsilon^{(v)}$, we have

$$e^{(k+1)} = (I - S(S^T AS)^{-1} S^T A)\varepsilon^{(v)} \tag{2.8}$$

$$= (I - S(S^T AS)^{-1} S^T A)G^v e^{(k)}. \tag{2.9}$$

By (2.9),

$$A^{1/2}e^{(k+1)} = (I - A^{1/2}S(S^T AS)^{-1}S^T A^{1/2})A^{1/2}G^v A^{-1/2}A^{1/2}e^{(k)}.$$

Note that

$$P_1 \equiv A^{1/2}S(S^T AS)^{-1}S^T A^{1/2} \tag{2.10}$$

is the orthogonal projector (with respect to the Euclidean inner product) onto the range of $A^{1/2}S$. Hence

$$A^{1/2}e^{(k+1)} = (I - P_1)A^{1/2}G^v A^{-1/2}A^{1/2}e^{(k)}, \tag{2.11}$$

whence

$$\|e^{(k+1)}\|_E \le \|I - P_1\|_2 \, \|G^{\upsilon}\|_E \, \|e^{(k)}\|_E$$
$$= \|G^{\upsilon}\|_E \, \|e^{(k)}\|_E. \tag{2.12}$$

Recalling that $G = I - Q^{-1}A$ for a basic iteration matrix $G$, we immediately have the following conclusion.

<u>THEOREM 2.1</u>. *If* $Q$ *is symmetric and* $\rho(G) < 1$, *then any sequence produced by Algorithm 1 with Galerkin approximation converges to the true solution* $u$.

*Proof.* By (2.12),

$$\begin{aligned}
\|e^{(k+1)}\|_E &\le \|G^{\upsilon}\|_E \, \|e^{(k)}\|_E \\
&\le \|G\|_E^{\upsilon} \, \|e^{(k)}\|_E \qquad (2.13) \\
&= \|I - Q^{-1}A\|_E^{\upsilon} \, \|e^{(k)}\|_E \\
&= \|I - A^{1/2}Q^{-1}A^{1/2}\|_2^{\upsilon} \, \|e^{(k)}\|_E \\
&= \|A^{1/2}GA^{-1/2}\|_2^{\upsilon} \, \|e^{(k)}\|_E \\
&= \rho^{\upsilon}(G) \, \|e^{(k)}\|_E.
\end{aligned}$$

QED.

<u>THEOREM 2.2</u>. *If we use the Gauss-Seidel method for the presmoother in Algorithm 1 with Galerkin approximation, then any sequence produced by Algorithm 1 is convergent to the true solution* $u$.

*Proof.* According to a theorem in Young [1971, p. 79],

$\|G\|_E < 1$ if and only if $Q^T + Q - A$ is SPD.

Noting (2.13) and applying the theorem to the Gauss-Seidel iteration matrix, we obtain Theorem 2.2.

QED.

Now by (1.5a) with $\tilde{e} = e^{(k+1)}$ and $e = \varepsilon^{(\upsilon)}$, we have

$$\|e^{(k+1)}\|_E = \min_{z \in Rg(S)} \|\varepsilon^{(\upsilon)} - z\|_E$$

$$\leq \|G^{\upsilon} e^{(k)} - z\|_E \qquad \forall z \in Rg(S). \qquad (2.14)$$

Inequality (2.14) suggests that we choose $Rg(S)$ to be the subspace spanned by $\{G^j e^{(k)}\}$. This is what we shall do in the next section.

## 3. Reduced Basis Technique

From (2.9) and Section 1, we know that $e^{(k+1)}$ is the orthogonal projection of $G^{\upsilon} e^{(k)}$ onto the $Rg(S)$. Hence the column vectors of the matrix $S$ can be replaced by any basis of the range $Rg(S)$. What matters is the subspace spanned by the column vectors of $S$. It is now clear that once we specify a subspace of $R^n$, Algorithm 1 is then completely defined.

From now on we shall denote one cycle of Algorithm 1 with Galerkin approximation by $RBM(G,S_R,\upsilon)$. Here RBM stands for the reduced basis method and $S_R$ is the reduced subspace. Consider the following choice of $S_R$, which was first proposed by Porsching [1990].

Pseudoresidual Based $RBM(G,S_R,\upsilon)$:

$$S_R \equiv \mathrm{span}\{\bar{\delta}^{(1)}, \bar{\delta}^{(2)}, \dots, \bar{\delta}^{(\upsilon)}\},$$

where

$$\bar{\delta}^{(i)} = w^{(j)} - u^{(k)}, \quad j = 1,2,\dots,\upsilon. \qquad (3.1)$$

Let $K^{(\upsilon)}(x,B) \equiv \mathrm{span}\{x, Bx, \dots, B^{\upsilon-1}x\}$, $x \in R^n$, $B \in R^{n \times n}$.

LEMMA 3.1. *Define* $\delta^{(k)} \equiv Gu^{(k)} + Q^{-1}b - u^{(k)}$. *Suppose that* $I - G$ *is invertible, then*

$$S_R = \mathrm{Span}\{\bar{\delta}^{(1)}, \dots, \bar{\delta}^{(\upsilon)}\}$$

$$= K^{(\upsilon)}(\delta^{(k)}, G) = K^{(\upsilon)}(\delta^{(k)}, I - G).$$

*Proof.* $\bar{\delta}^{(j)} = w^{(j)} - u + u - u^{(k)}$

$$= -G^j e^{(k)} + e^{(k)}$$

$$= (I - G^j)e^{(k)}. \qquad (3.2)$$

$$\delta^{(k)} = Gu^{(k)} + Q^{-1}b - u^{(k)}$$

$$= Gu^{(k)} + Q^{-1}Au - u^{(k)}$$

$$= (I - G)e^{(k)}. \qquad (3.3)$$

Hence

$$\bar{\delta}^{(j)} = (I - G^j)(I - G)^{-1}\delta^{(k)}$$

$$= (I - G)(I + \dots + G^{j-1})(I - G)^{-1}\delta^{(k)}$$

$$= (I + G + \dots + G^{j-1})\delta^{(k)}.$$

QED.

We remark that the condition $I - G$ is invertible holds for any completely consistent basic iterative method. This is true, for most commonly known basic iterative methods. $\delta^{(k)}$ is often called pseudoresidual vector. From now on, we shall denote the pseudoresidual-based $RBM(G,S_R,\upsilon)$ as $RRBM(G,S_R,\upsilon)$.

<u>THEOREM 3.1</u> [Chou and Porsching, 1988] *Given a* $RRBM(G,S_R,\upsilon)$ *with* $G = I - Q^{-1}A$ *defining a completely consistent presmoother. If* $Q$ *is SPD then*

*a)* $G$ *has real eigenvalues* $\{\mu_j\}$, $j = 1,2,\dots,n$, *which may be ordered as*

$$\mu_1 \leq \mu_2 \leq \dots \leq \mu_N < 1.$$

*b)* $\|e^{(k+1)}\|_E \leq \eta(\upsilon)\|e^{(k)}\|_E$, *where*

$$\eta(\upsilon) = 2\gamma^{\upsilon/2}/(1 + \gamma^{\upsilon}) < 1,$$

$$\gamma + (1 - \sqrt{1 - \sigma^2})/(1 + \sqrt{1 - \sigma^2}),$$

$$\sigma = (\mu_N - \mu_1)/(2 - \mu_N - \mu_1). \qquad (3.4)$$

*Proof.* Statement (a) follows easily by noting that

$$A^{1/2}GA^{-1/2} = I - A^{1/2}Q^{-1}A^{1/2}.$$

Statement (b) can be proved as follows. By (2.14) with $Rg(S) = S_R$,

$$\|e^{(k+1)}\|_E \le \|G^{\upsilon}e^{(k)} - \sum_{j=0}^{\upsilon-1} \alpha_j G^j \delta^{(k)}\|_E$$

$$= \|G^{\upsilon}e^{(k)} - \sum_{j=0}^{\upsilon-1} \alpha_j G^j (I - G)e^{(k)}\|_E$$

$$= \|p(G)e^{(k)}\|_E, \text{ for all polynomials } p(x) \text{ of } \deg \le \upsilon \text{ and } p(1) = 1.$$

$$\le \|A^{1/2}p(G)A^{-1/2}\|_2 \|e^{(k)}\|_E$$

$$= \rho(p(A^{1/2}GA^{-1/2}))\|e^{(k)}\|_E$$

$$= \rho(p(G))\|e^{(k)}\|_E.$$

Hence $\|e^{(k+1)}\|_E \le \min\limits_{\substack{\deg p \le \upsilon \\ p(1) = 1}} \max\limits_{\mu_1 \le \lambda \le \mu_N} |p(\lambda)| \, \|e^{(k)}\|_E$. Define

$\eta(\upsilon) \equiv \min\limits_{\substack{\deg p \le \upsilon \\ p(1)=1}} \max\limits_{\mu_1 \le \lambda \le \mu_N} |p(\lambda)|$. The conclusion of the theorem now follows from the well-known Chebyshev minimax theorem [Young, 1971, p. 302].

QED.

We now give a theorem that characterizes $u^{(k+1)}$ knowing $u^{(k)}$.

<u>THEOREM 3.2</u>. *Given a* $RRBM(G, S_R, \upsilon)$ *with initial iterate* $u^{(k)}$. *Then* $u^{(k+1)}$ *minimizes the quadratic functional* $F(w) \equiv \|w - u\|_E^2$ *over the linear manifold* $u^{(k)} + S_R$. *That is*

$$\|u^{(k+1)} - u\|_E \le \|w - u\|_E \quad \forall w \in u^{(k)} + K^{(\upsilon)}(\delta^{(k)}, I - G). \tag{3.5}$$

*Proof.* By (2.14),

$$\begin{aligned}
\|e^{(k+1)}\|_E &= \min_{z\in S_R} \|G^{v}e^{(k)} - z\|_E \\
&= \min_{z\in S_R} \|\varepsilon^{(v)} - z\|_E \\
&= \min_{z\in S_R} \|u - w^{(v)} - z\|_E \\
&= \min_{w\in W^{(v)}+S_R} \|w - u\|_E .
\end{aligned}$$

The theorem will be proved if we can show that

$$w^{(v)} + S_R = u^{(k)} + S_R. \tag{3.6}$$

Now

$$\begin{aligned}
w^{(v)} &= \sum_{j=1}^{v} (w^{(j)} - w^{(j-1)}) + w^{(0)} \\
&= \sum_{j=1}^{v} (Gw^{(j-1)} + Q^{-1}b - w^{(j-1)}) + u^{(k)} \qquad \text{(by 2.1a)} \\
&= \sum_{j=1}^{v} (I - G)(u - w^{(j-1)}) + u^{(k)} \\
&= \sum_{j=1}^{v} (I - G)G^{j-1}e^{(k)} + u^{(k)} \\
&= \sum_{j=1}^{v} (I - G)G^{j-1}(I - G)^{-1}\delta^{(k)} + u^{(k)} \\
&= \sum_{j=1}^{v} G^{j-1}\delta^{(k)} + u^{(k)}.
\end{aligned} \tag{3.7}$$

Hence $w^{(v)} - u^{(k)} \in S_R$.

QED.

<u>COROLLARY 3.2.1</u>. RRBM($G, S_R, v$) *with an initial iterate* $u^{(k)}$ *is equivalent to* $v$ *steps of conjugate gradient method with initial iterate* $u^{(k)}$, *provided that* $G = I - A$.

*Proof.* Note that

$$\begin{aligned} S_R &= \mathrm{Span}\{\delta^{(k)},(I - G)\delta^{(k)},\ldots,(I - G)^{\upsilon-1}\delta^{(k)}\} \\ &= \mathrm{Span}\{\delta^{(k)},A\delta^{(k)},\ldots,A^{\upsilon-1}\delta^{(k)}\} \\ &= \mathrm{Span}\{r^{(k)},Ar^{(k)},\ldots,A^{\upsilon-1}r^{(k)}\}, \quad r^{(k)} = b - Au^{(k)}. \end{aligned}$$

By (3.5),

$$\|u^{(k+1)} - u\|_E \le \|w - u\|_E \qquad \forall w \in u^{(k)} + K^{(\upsilon)}(r^{(k)},A).$$

QED.

Corollary 3.2.1 suggests a procedure for implementing $RRBM(G,S_R,\upsilon)$ based on Theorem 3.2. Recall that minimization of a quadratic functional induced by a SPD matrix $B$ over a k-plane can be achieved by minimizations of the quadratic functional along $k$ B-conjugate directions on the k-plane (see Hestenes, 1980, p. 101). Thus we can implement $RRBM(G,S_R,\upsilon)$ by the conjugate direction method and avoid explicit implementation of (2.1a), (2.1b), (2.2)-(2.4) altogether.

## 4. Implementation of $RRBM(G,S_R,\upsilon)$

Perform the following steps for $n = 1,\ldots,\upsilon - 1$.

$$w^{(0)} = u^{(k)},$$

$$w^{(n+1)} = w^{(n)} + \lambda_n p^{(n)},$$

$$p^{(n)} = \begin{cases} \delta^{(0)} & \text{if } n = 0 \\ \delta^{(n)} + \alpha_n p^{(n-1)}, & n = 1,2,\ldots,\upsilon - 1, \end{cases}$$

$$\alpha_n = -(\delta^{(n)},Ap^{(n-1)})/(p^{(n-1)},Ap^{(n-1)}),$$

$$\delta^{(n)} = Gw^{(n)} + Q^{-1}b - w^{(n)},$$

$$\begin{aligned} \lambda_n &= (p^{(n)},Q\delta^{(n)})/(p^{(n)},Ap^{(n)}), \\ &= (p^{(n)},r^{(n)})/(p^{(n)},Ap^{(n)}), \end{aligned}$$

$$r^{(n)} = b - Aw^{(n)},$$

$$G = I - Q^{-1}A. \tag{4.1}$$

Set $u^{(k+1)} = w^{(\upsilon)}$.

Here $(\cdot,\cdot)$ denote the Euclidean inner product. $\alpha_n$ is obtained by insisting that $(p^{(i)},Ap^{(j)}) = 0$, $i \neq j$. The $\lambda_n$ is obtained as follows.

Since $w^{(n+1)} = w^{(n)} + \lambda_n p^{(n)}$, it is not hard to see that

$$\delta^{(n+1)} + \lambda_n p^{(n)} = \delta^{(n)} + \lambda_n G p^{(n)}$$

and

$$r^{(n+1)} - r^{(n)} = -\lambda_n A p^{(n)}. \qquad (4.2)$$

Furthermore, $w^{(n+1)}$ is the minimizer along the direction $p^{(n)}$. Hence $(r^{(n+1)},p^{(n)}) = 0$. This together with (4.2) implies

$$\lambda_n = (r^{(n)},p^{(n)})/(p^{(n)},Ap^{(n)}).$$

## 5. Relationship between RRBM($G,S_R,\upsilon$), $G = I - Q^{-1}A$ and Preconditioned Conjugate Procedures

Let $(I - G)u = c$, where $c = Q^{-1}b$. Furthermore let $I - G$ be symmetrizable with a symmetrization matrix $\overline{W}$ [Hegaman and Young, 1981]. Define

$$\hat{A} \equiv \overline{W}(I - G)\overline{W}^{-1},$$
$$\hat{u} \equiv \overline{W}u,$$
$$\hat{b} \equiv \overline{W}c.$$

Then $\hat{A}\hat{u} = \hat{b}$ is called the preconditioned system with respect to $\overline{W}$. One can apply conjugate gradient method to this system. Then the following formulae are obtained [Hegeman and Young, 1981, p. 146].

$$u^{(0)} \text{ is arbitrary}$$
$$u^{(n+1)} = u^{(n)} + \lambda_n p^{(n)}, \quad n = 0,1,\ldots,$$
$$p^{(n)} = \begin{cases} \delta^{(0)}, & n = 0, \\ \delta^{(n)} + \alpha_n p^{(n-1)}, & n = 1,2,\ldots, \end{cases}$$
$$\alpha_n = -(\overline{W}\delta^{(n)},\overline{W}(I - G)p^{(n-1)})/(\overline{W}p^{(n-1)},\overline{W}(I - G)p^{(n-1)}), \quad n = 1,2,\ldots,$$
$$\lambda_n = (\overline{W}p^{(n)},\overline{W}\delta^{(n)})/(\overline{W}p^{(n)},\overline{W}(I - G)p^{(n)}), \quad n = 0,1,2,\ldots .$$

Here

$$\delta^{(n)} = Gu^{(n)} + c - u^{(n)}. \qquad (5.1)$$

**THEOREM 5.1**. *Let* RRBM($G,S_R,\upsilon$) *be such that* $G = I - Q^{-1}A$ *has a SPD splitting matrix* $Q$. *Then* RRBM($G,S_R,\upsilon$) *with an initial iterate* $\tilde{u}$ *is* *equivalent to the preconditioned conjugate gradient procedure (5.1), with* $\overline{W} = Q^{1/2}$, $u^{(0)} = \tilde{u}$ *and the upper limit of* $n = \upsilon$. *In particular,* $RRBM(G,S_R,\infty)$ *is equivalent to the preconditioned conjugate gradient procedure, provided that the same initial iterate is taken.*

*Proof.* Compare (5.1) with (4.1). (See Luenberger [1984, p. 245].)

QED.

We remark that if $Q$ is not SPD, the two methods are not equivalent. On the other hand, (4.1) is still applicable for nonsymmetrizable cases. For instance, Theorem 2.2 guarantees the convergence of $\bigcup_{\upsilon} RRBM(G,S_R,\upsilon)$ when $G$ is the iteration matrix of the Gauss-Seidel smoother.

## 6. Contraction Numbers in the Positive Real Case

In this section we derive generic contraction numbers for a class of additive correction methods based on orthogonal projection. The only assumption on the range of the projector is that it contains the residual. This generalizes the results of Section I (SPD case). However, the contraction number obtained here is not as sharp as the earlier one. The general approach below provides convergence results for restarted generalized conjugate gradient methods under a variety of conditions.

Consider the problem of determining $u \in R^n$ such that

$$Au = f, \qquad (6.1)$$

where $A \in R^{n\times n}$ is invertible, and $f \in R^n$.

Certain popular iterative methods for the solution of (6.1), such as generalized conjugate gradient (GCG) methods (Hageman and Young [1981, p. 339], Elman [1982], Saad and Schultz [1985], Vatsya [1988]) and multigrid methods (Hackbusch [1985], McCormick [1987]) incorporate into their overall solution strategies an additive correction algorithm of the following type: Given an approximation $u_0$ of $u$ and an $m$ dimensional subspace $S_m \subset R^n$:

1. Compute the residual

$$r_0 = f - Au_0. \tag{6.2}$$

2. Observe that the error $e_0 \equiv u - u_0$ satisfies

$$Ae_0 = r_0. \tag{6.3}$$

3. Compute an approximation $\tilde{e}_0 \in S_m$ of $e_0$.

4. Form a corrected approximation of $u$,

$$u_1 = u_0 + \tilde{e}_0. \tag{6.4}$$

The key to this additive correction phase is, as in Section 1, the method of determining $\tilde{e}_0$. In many instances (for example, GCG methods) this is done by an orthogonal projection $e_0$ onto $S_m$ with respect to a suitably defined inner product. Thus if $F \in R^{n \times n}$ is a given SPD operator, and we define the inner product

$$(\cdot,\cdot)_F = (\cdot,F\cdot),$$

with induced norm $\|\cdot\|_F$, then

$$\tilde{e}_0 = \Pi_{S_m} e_0. \tag{6.5}$$

where $\Pi_{S_m}$ is the orthogonal projector of $R^n$ onto $S_m$ with respect to $(\cdot,\cdot)_F$.

To obtain a more concrete representation of $\Pi_{S_m}$, we assume that $S_m = Rg(S)$ for some $S \in R^{n \times m}$. Then one easily has

$$\Pi_{S_m} = S(S^tFS)^{-1} S^tF. \tag{6.6}$$

Note that by (6.5) and (6.3),

$$\tilde{e}_0 = S(S^tFS)^{-1} S^tFA^{-1} r_0. \tag{6.7}$$

Hence, the prescription (6.5) is practical only if $F$ is of the form

$$F = EA \tag{6.8}$$

for some $E \in R^{n \times n}$. If $A$ is SPD, then an obvious and common choice of $E$ is the identity $I$ (i.e. $F = A$). Note that in Section 1 we also took $E = I$.

In terms of the element $u$ and $u_1$, equation (6.5) is equivalent to the condition

$$\|u - u_1\|_F \le \|v - u\|_F, \qquad (6.9)$$

for any element $v$ of the coset $u_0 + S_m$. Again, if $A$ is SPD and $F = A$, then $u_1$ is also the minimizer over $u_0 + S_m$ of the quadratic functional

$$\varphi(v) = (v,Av) - 2(v,f).$$

This is the usual condition required of the classical conjugate gradient iterates (see, for example Golub and VanLoan [1983, p. 362]) while condition (6.9) is used in the more general case (Hageman and Young [1981, p. 342]).

In this section we give conditions under which there is a contraction number $\gamma > 1$ such that if (6.5) defines the additive correction procedure and $e_1 = u = u_1$, then

$$\|e_1\|_F \le \gamma\|e_0\|_F. \qquad (6.10)$$

This establishes the convergence of any iterative method consisting solely of additive corrections satisfying (6.10). Moreover, given a more general iterative scheme which includes additive corrections for which (6.10) holds, convergence follows whenever the other phases of an iterative step do not increase the error. This is the case, for instance, with restarted versions of the conjugate gradient and certain GCG methods.

A matrix $B$ is positive real (PR) if the symmetric part of $B$ is SPD. With this in mind we state a theorem whose proof can be found in Chou and Porsching (1990). The theorem provides sufficient conditions for an estimate of the type (6.10).

THEOREM 6.1. *Let* $u_0$ *be given and let* $r_0$, $\tilde{e}_0$, *and* $u_1$, *be given by (6.2), (6.5) and (6.4). If* $r_0 \in S_m$ *and* $FA^{-1}$ *is PR, then the error* $e_i = u - u_i$, $i = 0,1$ *satisfy*

$$\|e_1\|_F \le \gamma(F^{1/2} A^{-1} F^{-1/2})\|e_0\|_F,$$

*where* $\gamma(\cdot)$ *is the contraction number defined by*

$$\gamma(B) = [1 - \lambda^2_{min}(B + B^t)/(4\lambda_{max}(B^tB))]^{1/2} < 1.$$

Now we consider the application of Theorem 6.1 to some specific types of systems (6.1). Note that in terms of the matrix $E$ defined by (6.8), the hypotheses of the theorem require that $E$ be PR and $EA$ SPD.

If A itself is SPD, then as previously observed, we can take $F = A$. In this case the contraction number of Theorem 6.1 is

$$\gamma(A^{-1}) = \left[1 - \frac{\lambda^2_{\min}(A^{-1})}{\lambda^2_{\max}(A^{-1})}\right]^{1/2} = \left[1 - \frac{1}{k_2^2(A)}\right]^{1/2} .$$

where $k_2(A)$ denotes the spectral number of A. This should be compared with the contraction number $(k_2(A) - 1)/(k_2(A) + 1)$ previously obtained by Chou and Porsching (1989) for this special case. (See also Section 1.) It is easy to see that $\gamma(A^{-1}) \geq (k_2(A) - 1)/(k_2(A) + 1)$, with equality if and only if $A = I$.

Next we consider the case when A is PR. Since $A^t$ is then also PR, we can take $F = A^tA$. The contraction number is $\gamma((A^tA)^{1/2} A^{-1} (A^tA)^{-1/2})$, but this can be simplified by applying the following lemma (see Chou and Porsching [1990]).

LEMMA 6.2. *Let* $A, P \in R^{n\times n}$ *be respectively nonsingular and SPD. Then for any real number* $\alpha$,

$$\|I - \alpha(A^tP^{-1}A)^{1/2} A^{-1}P(A^tP^{-1}A)^{-1/2}\|_2 = \|I - \alpha P^{1/2}A^{-1}P^{1/2}\|_2.$$

If A is PR, we apply the lemma with $P = I$. The result is that the contraction number $\gamma((A^tA)^{1/2} A^{-1}(A^tA)^{-1/2})$ may be replaced by $\gamma(A^{-1})$.

Finally, we turn to the generalized conjugate acceleration procedures as presented in Hageman and Young [1981]. Let $Q \in R^{n\times n}$ be a nonsingular splitting or preconditioning matrix. Then (6.1) can be written as

$$(I - G)u = b, \tag{6.11}$$

where $G = I - Q^{-1}A$ and $b = Q^{-1}f$. In the context of the basic iterative method

$$v_{k+1} = Gv_k + b \quad (v_0 = u_0), \quad k = 0,1,...., \tag{6.12}$$

the quantity $r_0 = b - (I - G)u_0$ is known as a "pseudoresidual". Note that $r_0 = v_1 - v_0$.

We assume (as do Hageman and Young [1981, p. 341]) that there is a matrix Z such that $Z(I - G)$ is SPD. If $S_m$ is the m dimensional (Krylov) subspace spanned by the elements $r_0$, $(I - G)r_0$,...., and if $F = Z(I - G)$, then our additive correction method is equivalent to the generalized conjugate gradient method termed ORTHODIR by Hageman and Young. If $Z$ is PR, then the method becomes the so-called ORTHOMIN method. It follows immediately from the definitions of $S_m$ and $F$ that Theorem 6.1 applies to ORTHOMIN with the contraction number $\gamma((Z(I - G))^{-1/2} Z(Z(I - G))^{-1/2})$.

The GCW generalized conjugate gradient method (after Concus and Goulb [1976] and Widlund [1978]) is obtained from (6.11) by choosing $Q = 1/2(A + A^t)$ under the assumption that $A$ is PR. In this case we can take $F = A^tQ^{-1}A$ and Theorem 6.1 holds with respect to the contraction number $\gamma((A^tQ^{-1}A)^{1/2} A^{-1}Q(A^tQ^{-1}A)^{-1/2})$. By applying Lemma 6.2 with $P = Q$, we see that this contraction number can be replaced by $\gamma(Q^{1/2}A^{-1}Q^{1/2})$.

## 7. Hybrid Difference Methods

In this section we study hybrid difference methods for the linear convection equation

$$u_t + A(x)u_x = 0, \quad t \geq 0, \qquad (7.1)$$

subject to the initial condition,

$$u(x,0) = \psi(x).$$

Here $u: R \to R^n$, $A: R \to R^{n \times n}$, $\psi: R \to R^n$ are smooth functions. The matrix function $A$ is required to be symmetric, and there exists a constant $\mu$ such that $uI \geq A > 0$. (The ordering here is that of symmetric matrices.) By a hybrid diffference method we mean a method that is obtained by forming weighted combinations of difference quotients defining two consistent methods. While there are many ways to select weights in the blending process, we shall concentrate on a principle suggested by Porsching [1989] for the equation (7.1) in the case of $n = 1$. His idea is to examine the continuous problem (7.1) to see if a conservation of energy can be found. If that can be found, then one tries to create a blending process so that a similar energy conservation holds for the corresponding discrete case. Thus we start with problem (7.1) and derive a conservation of weighted energy for it. Without loss of generality, we shall assume the data $A(x)$, $\psi(x)$ given in (7.1) is spatially periodic with period $\ell$. The energy associated with (7.1) is defined as

$$I^2(t) \equiv \int_0^{\ell} (u(x,t),u(x,t))dx,$$

where $(\cdot,\cdot)$ denotes the Euclidean inner product. Note that the energy so defined is nothing but the $L^2$-norm of the function $u(\cdot,t)$ if we define

$$\langle v,w\rangle \equiv \int_0^{\ell} (v,w)dx \quad \text{for any} \quad v,w \in L^2(0,\ell).$$

It is well known that spatially periodic symmetric hyperbolic system has a unique periodic solution (see Kreiss [1989, p. 100]). Hence using the periodicity of $u$ and $A$, it is easy to see that

$$\langle u,A_x u\rangle = -2\langle u,Au_x\rangle.$$

Thus

$$\begin{aligned} \frac{d}{dt}\langle u,u\rangle &= 2\langle u_t,u\rangle \\ &= -2\langle Au_x,u\rangle \\ &= \langle u,A_x u\rangle. \end{aligned}$$

Hence $\frac{d}{dt} I^2(t) \leq \|A_x\|_\infty \; I^2(t)$, which implies

$$I^2(t) \leq \exp(\|A_x\|_\infty t)I^2(0).$$

We see that $I^2(t)$ is bounded if we are interested in seeking solution over a finite interval $[0,T]$. In general, the energy is not conserved. However, the weighted energy

$$J^2(t) \equiv \int_0^{\ell} (u(x,t),A^{-1}(x) \; u(x,t))dx$$

is conserved. In fact

$$\frac{d}{dt} J^2(t) = \frac{d}{dt} \langle u, A^{-1}u \rangle = \langle u_t, A^{-1}u \rangle + \langle u, A^{-1}u_t \rangle \quad ,$$

$$= 2\langle u_t, A^{-1}u \rangle$$

$$= -2\langle Au_x, A^{-1}u \rangle$$

$$= 0.$$

It is the discrete analog of the weighted energy that we shall base our weighted selection principle on. Let the $x - t$ plane be covered by a rectangular mesh with uniform $x$ and $t$ spacing $h$ and $\tau$. Let $\upsilon$ be a mesh function whose values are in $R^n$ and whose value at the mesh $(jh, m\tau)$ is denoted by $\upsilon_j(m)$. When no confusion can arise we shall suppress the dependence on $j$ and/or $m$. For any such function we define the common x-directional differences:

$$S_x^{\pm}\upsilon = \upsilon_{j+1},$$

$$\Delta_x \upsilon = (\upsilon_{j+1} - \upsilon_j)/h,$$

$$\nabla_x \upsilon = (\upsilon_j - \upsilon_{j-1})/h,$$

$$\overline{\Delta}_x \upsilon = (\upsilon_{j+1} - \upsilon_{j-1})/2h,$$

$$\delta_x^2 \upsilon = (\upsilon_{j+1} - 2\upsilon_j + \upsilon_{j+1})/h^2,$$

as well as analogous differences in the $\tau$-direction. The following identities hold for any mesh functions $u$ and $\upsilon$:

$$\nabla_x(u, \Delta_x \upsilon) = (u, \delta_x^2 \upsilon) + (\nabla_x u, \nabla_x \upsilon), \tag{7.2}$$

$$\nabla_x(u, \upsilon) = (u, \nabla_x \upsilon) + (S_x^- \upsilon, \nabla_x \upsilon), \tag{7.3}$$

$$2(\upsilon, \nabla_x \upsilon) = \nabla_x(\upsilon, \upsilon) + h(\nabla_x \upsilon, \nabla_x \upsilon), \tag{7.4}$$

and

$$\Delta_t(u, \upsilon) = (u, \Delta_t \upsilon) + (\Delta_t u, S_t' \upsilon), \tag{7.5a}$$

$$2(B^{-1}\upsilon, \Delta_t \upsilon) = \Delta_t(B^{-1}\upsilon, \upsilon) - \tau(B^{-1}\Delta_t \upsilon, \Delta_t \upsilon), \tag{7.5b}$$

where $B = B(x) \in R^{n \times n}$ is symmetric.

These identities can be proven first for the case $n = 1$ and then for the general $n$ by applying the results to each component. (7.5b) follows from (7.5a) by setting $u = B^{-1}v$.

Now consider the following hybrid difference equation for (7.1):

$$\Delta_t v + \mathcal{L}[(1 - \theta)\nabla_x v + \theta\overline{\Delta}_x v] = 0, \qquad (7.6)$$

where $\mathcal{L}_j(m) = \mathcal{L}_j = A(jh)$ and $\theta$ is a scalar mesh function of weights. To be consistent with the continuous problem, we assume that $v$ is $L$-periodic in $j$, i.e. $v_{j+L}(m) = v_j(m)$. Equation (7.6) can be rewritten as

$$\Delta_t v + \mathcal{L}(\nabla_x v + \frac{1}{2}\theta h\, \delta_x^2 v) = 0, \qquad (7.7)$$

which clearly reveals its antidiffusive nature when $\theta > 0$. In the next few steps we shall transform (7.7) into a more amenable form from which the discrete energy can be singled out. The idea is to treat all new terms not present in the continuous case as parts of the "source" term. By source term we mean the term that tends to zero as $h$ and $\tau$ to zero.

From (7.7) we have

$$(\mathcal{L}^{-1}v, \Delta_t v) + (\mathcal{L}^{-1}v, \mathcal{L}[\nabla_x v + \frac{1}{2}\theta h\, \delta_x^2 v]) = 0.$$

Using (7.2) to transform the second term, we get

$$2(\mathcal{L}^{-1}v, \Delta_t v) + 2(v, \nabla_x v) + \nabla_x(\theta h v, \Delta_x v) - (\nabla_x(\theta h v), \nabla_x v) = 0. \qquad (7.8)$$

Next we use (7.5) and (7.4) to transform the first two terms of (7.8) and the identity $\nabla_x(\theta v) = (\nabla_x\theta)v + S_x^-\theta(\nabla_x v)$ to transform the fourth term. Upon rearranging, we have

$$\Delta_t(\mathcal{L}^{-1}v, v) + \nabla_x[(v, v) + (\theta h v, \Delta_x v)]$$

$$= \tau(\mathcal{L}^{-1}\Delta_t v, \Delta_t v) + h(\nabla_x v, S_x^-\theta(\nabla_x v)) - h(\nabla_x v, \nabla_x v) + h(\nabla_x v, (\nabla_x\theta)v). \qquad (7.9)$$

Adding $h\gamma(\nabla_x v, \nabla_x v)$ to both sides of (7.9), where $\gamma$ is a constant, we obtain

$$\Delta_t(\mathcal{L}^{-1}v, v) + \nabla_x[(v, v) + (\theta h v, \Delta_x v)] + h\gamma(\nabla_x v, \nabla_x v)$$

$$= \tau(\mathcal{L}^{-1}\Delta_t v, \Delta_t v) - h(1 - \gamma - S_x^-\theta)(\nabla_x v, \Delta_x v) + h(\nabla_x\theta, v\nabla_x v). \qquad (7.10)$$

Letting

$$b = \nabla_x \theta \qquad (7.11)$$

and

$$Q = \tau(\mathcal{L}^{-1}\Delta_t v, \Delta_t v) - h(1 - \gamma - S_x^- \theta)(\nabla_x v, \nabla_x v), \qquad (7.12)$$

we see that

$$\Delta_t(\mathcal{L}^{-1} v, v) + \nabla_x[(v,v) + (\theta h v, \Delta_x v)] + h\gamma(\nabla_x v, \nabla_x v) = Q + h(b, v\nabla_x v). \qquad (7.13)$$

If we define the weighted energy of the method (7.6) as

$$\tilde{J}^2(m) = \sum_{j=0}^{L-1} (v_j(m), \mathcal{L}_j^{-1} v_j(m))h,$$

then we can use (7.13) to bound the energy

$$\tilde{I}^2(m) = \sum_{j=0}^{L-1} (v_j(m), v_j(m))h,$$

provided that we impose proper conditions.

In the scalar case $n = 1$ of (7.1), Porsching [1989] has porposed two different strategies: global weight and local weight selection procedures.

However, we have only been able to generalize the global weight procedure to higher dimensional cases at this moment. We describe how this can be done. Assume that the scalar weight function $\theta$ is independent of the space idex $j$. In this case $b = 0$, and if we choose $\gamma = 0$, then (7.13) reduces to

$$\Delta_t(\mathcal{L}^{-1} v, v) + \nabla_x[(v,v) + (\theta h v, \Delta_x v)] = Q, \qquad (7.14)$$

where

$$Q = \tau(\mathcal{L}^{-1}\Delta_t v, \Delta_t v) - h(1 - \theta)(\nabla_x v, \nabla_x v). \qquad (7.15)$$

If we multiply (7.14) by $h\tau$, sum over $0 \le j \le L - 1$, $0 \le k \le m$, use the periodicity of $v$, and note that

$$\sum_{j,k} \Delta_t(\mathcal{L}^{-1} v, v)h\tau = \tilde{J}^2(m + 1) - \tilde{J}^2(0),$$

we see that

$$\tilde{J}^2(m + 1) = \tilde{J}^2(0) + \sum_{j,k} Qh\tau. \tag{7.16}$$

Hence the weighted energy is conserved if

$$\sum_{j,k} Qh\tau = 0. \tag{7.17}$$

To enforce (7.17) we observe

$$\frac{Q}{h} = (\lambda[\nabla_x v + \frac{1}{2}\theta h\delta_x^2 v],\nabla_x v + \frac{1}{2}\theta h\delta_x^2 v) - ((1 - \theta)\nabla_x v,\nabla_x v), \tag{7.18}$$

where $\lambda \equiv \tau\mathcal{L}/h$ is the mesh matrix function of "Courant numbers". Thus

$$\sum_j Qh\tau = (D\theta^2 + E\theta + F)h^2\tau,$$

where

$$D = \frac{h^2}{4} \sum_j (\delta_x^2 v,\lambda\delta_x^2 v),$$

$$E = \sum_j (\lambda h\delta_x^2 v,\nabla_x v) + (\nabla_x v,\nabla_x v), \tag{7.19}$$

$$F = \sum_j (\lambda\nabla_x v,\nabla_x v) - (\nabla_x v,\nabla_x v).$$

Hence (7.17) follows if $\theta$ is a real root of the quadratic function

$$q_2(\theta) \equiv D\theta^2 + E\theta + F. \tag{7.20}$$

If the "Courant condition" $\rho(\lambda)$ = the spectral radius of $\tau\mathcal{L}/h \le 1$ holds, then

$$q_2(0) = \sum_j (\lambda\nabla_x v,\nabla_x v) - (\nabla_x v,\nabla_x v) \le 0,$$

and

$$q_2(1) = \sum_j (\lambda^{1/2}\,\overline{\Delta}_x v,\lambda^{1/2}\overline{\Delta}_x v) \ge 0.$$

Hence (7.20) has a root in the interval $[0,1]$.

In the derivation so far, we have assumed the scalar weight function $\theta$ depends on the time-spacing, but not on the spatial-spacing. From now on, we further assume that $\theta$ is independent of the time index $k$ (i.e. $\theta$ = constant). In general, (7.17) can no longer hold. However, if we write (7.18) in the form

$$\frac{Q}{h} = (\lambda[(1 - \frac{1}{2}\theta)\nabla_x \upsilon + \frac{1}{2}\theta\Delta_x \upsilon],[(1 - \frac{1}{2}\theta)\nabla_x \upsilon + \frac{1}{2}\theta\Delta_x \upsilon]) - ((1 - \theta)\nabla_x \upsilon, \nabla_x \upsilon),$$

let $\Lambda = \mu\tau/h$, and use $2ab \le \varepsilon^2 a^2 + \varepsilon^{-2} b^2$, then we do have

$$\frac{Q}{h} \le \Lambda\{\|(1 - \frac{1}{2}\theta)\nabla_x \upsilon)\|^2 + 2\|(1 - \frac{1}{2}\theta)\nabla_x \upsilon\|\|\frac{1}{2}\theta\Delta_x \upsilon\| + \|\frac{1}{2}\theta\Delta_x \upsilon\|^2\}$$
$$- (1 - \theta)\|\nabla_x \upsilon\|^2$$
$$\le \Lambda\{(1 - \frac{1}{2}\theta)^2 (1 + \varepsilon^2)\|\nabla_x \upsilon\|^2 + \frac{1}{4}\theta^2(1 + \varepsilon^{-2})\|\nabla_x \upsilon\|^2\} - (1 - \theta)\|\nabla_x \upsilon\|^2,$$

where $\|\cdot\|$ denotes the Euclidean norm. Thus

$$\sum_j Qh\tau \le q_1(\theta) \sum_j \|\nabla_x \upsilon\|^2 h^2 \tau,$$

where

$$q_1(\theta) \equiv \frac{\Delta}{4}(2 + \varepsilon + \varepsilon^{-2})\theta^2 + [1 - (1 + \varepsilon^2)\Lambda]\theta + (1 + \varepsilon^2)\Lambda - 1. \qquad (7.21)$$

Note that if $0 < \Lambda \le \dfrac{1}{1 + \varepsilon^2}$, then

$$q_1(0) = (1 + \varepsilon^2)\Lambda - 1 \le 0$$

and

$$q_1(1) = \frac{\Delta}{4}(2 + \varepsilon^2 + \varepsilon^{-2}) > 0.$$

Thus there is a root in $[0,1]$.

We summarize our findings as follows

<u>THEOREM 7.1</u>. *If* $\Lambda = \mu\tau/h \le 1/(1 + \varepsilon^2)$ *and if* $\theta$ *is a positive root of the equation* $q_1(\theta) = 0$ *in (7.21) for any* $\varepsilon$, *then* $\theta \in [0,1)$ *and* $\tilde{J}(m) \le \tilde{J}(0)$, $m \ge 0$.

<u>THEOREM 7.2</u>. *If* $\frac{\tau}{h}\rho(A) \le 1$ *and* $\theta(m)$ *is the root of*

$$D(m)\theta^2 + E(m)\theta + F(m) = 0,$$

*lying in* $[0,1]$, *where*

$$D(m) = \frac{h^2}{4} \sum_j (\delta_x^2\, \upsilon_j(m), \lambda_j(m) \delta_x^2\, \upsilon_j(m)),$$

$$E(m) = \sum_j (\lambda_j(m) h \delta_x^2\, \upsilon_j(m), \nabla_x \upsilon_j(m)) + (\nabla_x \upsilon_j(m), \nabla_x \upsilon_j(m)),$$

$$F(m) = \sum_j (\lambda_j(m) \nabla_x\, \upsilon_j(m), \nabla_x\, \upsilon_j(m)) - (\nabla_x\, \upsilon_j(m), \nabla_x\, \upsilon_j(m)),$$

*then the weighted energy of (7.7) is conserved, i.e.,*

$$\tilde{J}(m) = \tilde{J}(0) \quad \textit{for all} \quad m \geq 0.$$

## ACKNOWLEDGMENT

The author is indebted to Professor Porsching for many helpful discussions and for drawing his attention to the multigrid methods.

## REFERENCES

1. Braess, D. [1981]. The contraction number of a multigrid method for solving the Poisson equation. *Numer. Math.* **37**, pp. 387-404.

2. Chou, S. H. and Porsching, T. A. [1989]. A note on contraction numbers for additive correction methods. *Appl. Math. Letters* **2**, pp. 83-86.

3. Chou, S. H. and Porsching, T. A. [1990]. Contraction numbers for additive correction methods. *Linear Algebra and Its Applications*, to appear.

4. Chou, S. H. and Porsching, T. A. [1990]. Reduced additive correction procedures. In preparation.

5. Concus, P. and Golub, G. H. [1976]. A generalized conjugate gradient method for nonsymmetric systems of linear equations. Rep. Stan-Cs-76-646, Computer Science Department, Stanford University.

6. Elman, H. C. [1982]. Iterative methods for large sparse nonsymmetric systems of linear equations. Ph.D. Thesis, Computer Science Department, Yale University.

7. Golub, G. H. and Van Loan, C. F. [1983]. *Matrix Computations.* Baltimore, Maryland: Johns Hopkins University Press.

8. Hackbusch, W. [1985]. *Multi-Grid Methods and Applications.* Berlin: Springer-Verlag.

9. Hageman, L. A., and Young, D. M. [1981]. *Applied Iterative Methods.* New York, NY: Academic Press.

10. Hall, C. A., Porsching, T. A. and Raymond, M. [1989]. Fast algorithm development for large-eddy simulation of circular-jet turbulence. Technical Report ICMA-89-133, Department of Mathematics, University of Pittsburgh.

11. Hestenes, M. R. [1980]. *Conjugate-Direction Methods in Optimization.* Berlin: Springer-Verlag.

12. Kreiss, H-O and Lorenz, J. [1989]. *Initial-Boundary Value Problems and Navier-Stokes Equations.* New York: Academic Press.

13. Luenberger, D. [1984]. *Introduction to Linear and Nonlinear Programming.* Reading, Mass.: Addison-Wesley.

14. McCormick, S. (ed.) [1987]. *Multigrid Methods.* Frontier of Applied Mathematics Series No. 3, Philadelphia: SIAM.

15. Saad, Y. [1981]. Krylov subspace methods for solving large unsymmetric linear systems. *Math. Comp.* **37**, pp. 105-126.

16. Saad, Y. and Schultz, M. H. [1985]. Conjugate gradient-like algorithms for solving nonsymmetric linear systems. *Math. Comp.* 44, pp. 417-424.

17. Vatysa, S. R. [1988]. Convergence of conjugate residual-like methods for solving linear equations. *SIAM J. Numer. Anal.* **25**, pp. 957-964.

18. Widlund, O. [1978]. A Lanzcos method for a class of nonsymmetric systems of linear equations. *SIAM J. Numer. Anal.* **15**, pp. 801-812.

19. Young, D. M. [1971]. *Iterative Solution of Large Linear Systems.* New York: Academic Press.

## DOCUMENTATION

The following reports were supported in part by this research program:

1. Chou, S. H. and Porsching, T. A. [1989]. A note on contraction numbers for additive correction methods. *Applied Math. Letters 2*, pp. 83-96.

2. Chou, S. H. and Porsching, T. A. [1990]. Contraction numbers for additive correction methods. *Linear Algebra and Its Applications*, to appear (partial content of BGSU Technical Report 89-03).

3. Chou, S. H. and Porshing, T. A. [1990]. Reduced additive correction procedures. In preparation.

4. Chou, S. H. [1987]. On choosing reduced basis subspace arising from multigrid methods. Technical Report 87-23, BGSU.